ORIO (Online Resource for Integrative Omics): a web-based platform for rapid integration of next generation sequencing data

5678–5690 *Nucleic Acids Research, 2017, Vol. 45, No. 10* *Published online 11 April 2017*
doi: 10.1093/nar/gkx270

ORIO (Online Resource for Integrative Omics): a web-based platform for rapid integration of next generation sequencing data

Christopher A. Lavender[1,†], Andrew J. Shapiro[2,†], Adam B. Burkholder[1], Brian D. Bennett[1], Karen Adelman[3,4] and David C. Fargo[5,*]

[1]Integrative Bioinformatics, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709, USA, [2]Program Operations Branch, Division of the National Toxicology Program, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709, USA, [3]Epigenetics and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709, USA, [4]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA and [5]Office of Scientific Computing, Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC 27709, USA

Received January 10, 2017; Revised April 03, 2017; Editorial Decision April 04, 2017; Accepted April 05, 2017

ABSTRACT

Established and emerging next generation sequencing (NGS)-based technologies allow for genome-wide interrogation of diverse biological processes. However, accessibility of NGS data remains a problem, and few user-friendly resources exist for integrative analysis of NGS data from different sources and experimental techniques. Here, we present Online Resource for Integrative Omics (ORIO; https://orio.niehs.nih.gov/), a web-based resource with an intuitive user interface for rapid analysis and integration of NGS data. To use ORIO, the user specifies NGS data of interest along with a list of genomic coordinates. Genomic coordinates may be biologically relevant features from a variety of sources, such as ChIP-seq peaks for a given protein or transcription start sites from known gene models. ORIO first iteratively finds read coverage values at each genomic feature for each NGS dataset. Data are then integrated using clustering-based approaches, giving hierarchical relationships across NGS datasets and separating individual genomic features into groups. In focusing its analysis on read coverage, ORIO makes limited assumptions about the analyzed data; this allows the tool to be applied across data from a variety of experiments and techniques. Results from analysis are presented in dynamic displays alongside user-controlled statistical tests, supporting rapid statistical validation of observed results. We emphasize the versatility of ORIO through diverse examples, ranging from NGS data quality control to characterization of enhancer regions and integration of gene expression information. Easily accessible on a public web server, we anticipate wide use of ORIO in genome-wide investigations by life scientists.

INTRODUCTION

With the advent of next generation sequencing (NGS) (1), a wide diversity of techniques for whole-genome characterization of biological processes has emerged. These techniques allow for interrogation of genetic sequence (DNA-seq), DNA accessibility (DNase-seq and ATAC-seq) (2,3), DNA-protein interactions (ChIP-seq) (4) and expression profiles (RNA-seq) (5), among other biological properties. Though valuable on their own, integration of these approaches provides a fuller picture of highly coordinated biological processes, such as gene regulation (6,7). Despite these advances, integrative analysis of NGS data remains inaccessible to many life scientists. Most existing tools for NGS data require specialized computational expertise that to-date has not been a core component of biology training. Further, accessible data integration tools primarily focus on visualization of data at a single locus (8,9), limiting genome-wide analyses.

To provide a platform for large-scale NGS data integration that empowers life scientists, we developed ORIO (Online Resource for Integrative Omics), a web-based tool for rapid analysis of NGS datasets (Figure 1). An ORIO anal-

*To whom correspondence should be addressed. Tel: +1 919 541 0762; Email: fargod@niehs.nih.gov
†These authors contributed equally to this work as first authors.

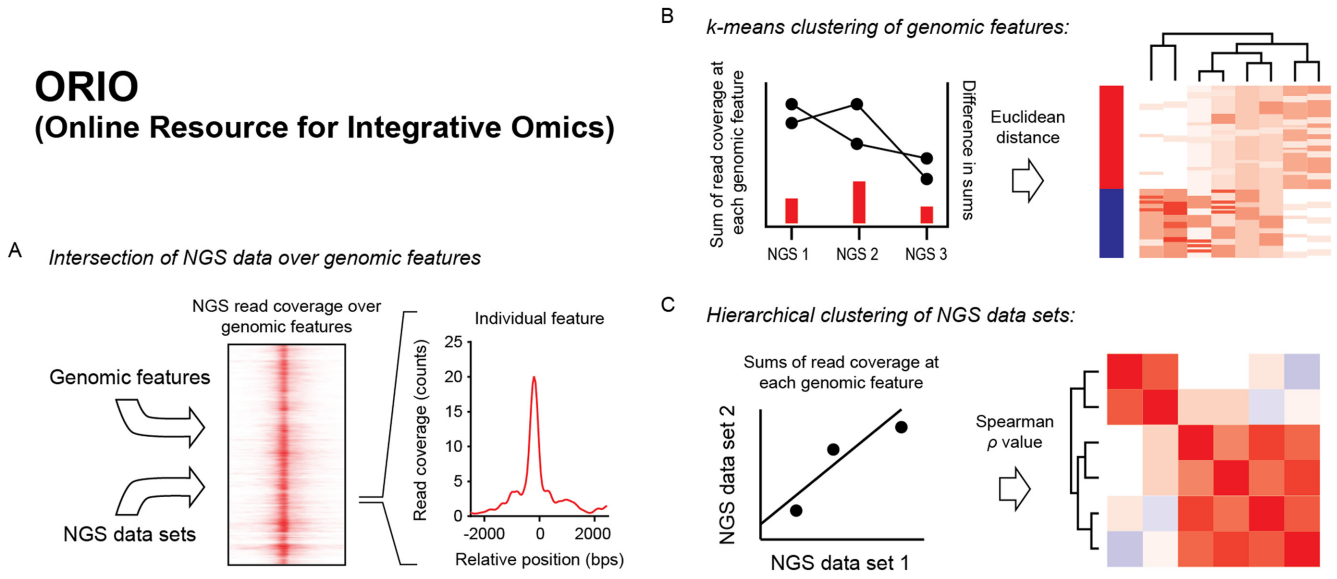

Figure 1. Schematic of analysis by ORIO. (**A**) Intersection of NGS data over genomic features. ORIO first finds read coverage values at each genomic feature for each NGS dataset in an analysis. Read coverage value are determined for genomic windows anchored on feature positions. (**B**) *k*-means clustering of genomic features. Read coverage values are used to cluster genomic features by *k*-means. To perform clustering, read coverage values for each NGS dataset are concatenated to make a 1D vector for each feature. The Euclidean distances between these vectors are then used in *k*-means clustering. (**C**) Hierarchical clustering of NGS datasets. Pairwise correlation values considering read coverage values at each feature are used as a distance metric in hierarchical clustering of datasets.

ysis begins with the user selecting NGS datasets of interest and specifying a list of loci as genomic coordinates. These coordinates can correspond to biologically relevant genomic features, such as transcription start sites or genomic locations of ChIP-seq peaks. ORIO first iteratively calculates the read coverage at genomic features for each NGS dataset (Figure 1A). ORIO provides dynamic display options to investigate these read coverage values, including heatmaps with extensive options for rank ordering. To support discovery-based investigation of these coverage values, ORIO then performs clustering across datasets, grouping genomic features into informative groups (Figure 1B) and finding hierarchical relationships across NGS datasets (Figure 1C). Clustering can have functional implications important to discovery, implying coordinated regulation or direct interaction.

ORIO is implemented in a modern web framework that organizes data and analysis results. All features are accessible using its web interface; users may upload data, set up analyses and view results. ORIO also hosts 4506 human and mouse datasets from the ENCODE research project, providing a point of access for life scientists to contextualize their own data within a rigorously controlled dataset. Statistical tests are also implemented next to dynamic displays of analysis results, allowing transitions from discovery to hypothesis-based inquiry over iterative analysis. ORIO was consciously designed to make minimal assumptions about data during analysis, allowing its applications to a variety of experiment types and study designs. We present ORIO alongside several example analyses to illustrate its versatility. These examples range from quality control of a target dataset to integration of NGS data with gene expression information and genome-wide characterization of enhancer regions.

MATERIALS AND METHODS

ORIO analysis

ORIO anchors its analysis of NGS datasets on a user-defined feature list of genomic coordinates. The first step of an ORIO analysis is selection of NGS datasets (up to 500 individual datasets) and a feature list from public or user-uploaded options. Feature lists of genomic coordinates are accepted in Browser Extensible Data (BED) format (8), facilitating its use with other bioinformatics tools. BED files may contain up to 500 000 features, allowing for comprehensive analysis of most genome-wide phenomena. ORIO accepts BED files with three or more columns. ORIO takes strand information from a BED file if available and uses it to orient coverage calculations for individual features.

ORIO first iteratively finds read coverage for NGS datasets at each genomic feature. Read coverage is determined within user-specified windows anchored at each genomic feature. By default, the window originates 2500 nts upstream of the center of a given feature and extends downstream by fifty 100-nt non-overlapping bins; the total default window size is 5000 nts. The window start, bin number and bin size may all be changed by the user. The user may also specify strand for both the genomic feature list and NGS datasets. If specified, values will reflect same-strand coverage. Observation of strand-specificity is especially important in the context of gene models where only a single strand is transcribed and where the genomic environment may be distinct upstream and downstream of a given feature (10). Read coverage values are compiled into a matrix

file for each NGS dataset. Matrix files are accessed internally by ORIO and may be downloaded by the user through the web interface.

ORIO performs correlative analysis considering read coverage values. A single total read coverage value is found for each genomic feature by adding the coverage from NGS datasets across all bins in a genomic window. Clustering methods are then applied to total read coverage values considering both NGS datasets and individual genomic features. To group NGS datasets, the Spearman correlation value for each pair of datasets is found by comparing feature coverage values. The use of rank-order Spearman correlation values mitigates the need of normalization of read coverage values between datasets. NGS datasets are then grouped by hierarchical clustering considering $1 - \text{abs}(\rho)$ as a distance metric. To group genomic features, a 1D vector is first found for each feature by concatenating coverage totals for each NGS dataset. These values are normalized by variance, and the Euclidean distance between these vectors is used to group genomic features by k-means clustering. k-means clustering is iteratively performed for k-values 2 through 10, and clustering results for any of these k-values may be selected and accessed by the user in the interface.

An alternative analytical approach is used if a sort vector is specified at the onset of analysis. A sort vector is a flexible data format where a single value is specified for each genomic feature. A sort vector is useful for integrating information into an ORIO analysis that may not be easily inferred from read coverage and may require additional assumptions in analysis. For instance, we present an example in this work where gene expression information is incorporated using a sort vector. When a sort vector is specified, Spearman correlations are determined between sort vector values and calculated NGS read coverage for each window bin. Correlation values across each window are then concatenated, resulting in a single data vector for each NGS dataset. Hierarchical clustering is then performed considering Euclidean distance between vectors as a distance metric. k-means clustering of genomic features is performed in the same way as in analyses without a sort vector.

ORIO accepts coverage information from NGS datasets in bigWig format. Consortia such as ENCODE report read coverage in this format. Increasingly, other groups are following the same practice, allowing other users access to coverage values consistent with alignment and read processing procedures of the original authors. Coverage values may be derived from actual reads or from estimated fragments, as appropriate for each individual dataset, and may be readily derived from alignment files using publically available software such as BEDTools (11). ORIO achieves fast runtimes by parallelizing read coverage calculations and performing computationally intensive calculations in C utilities. Coverage values are calculated using bigWigAverageOverBed from the UCSC Kent Tools utilities package (12). ORIO runtimes range from seconds to hours, depending on the number of datasets analyzed and the length of the genomic feature list. Benchmarking results are given in Table 1.

ORIO analysis is performed using the ORIO Python package available at https://github.com/niehs/orio/. For clustering and other applications, ORIO uses NumPy and SciPy (13).

ORIO web framework

ORIO is hosted on a public server. ORIO uses the Python-based Django web framework (14) that allows for robust management of user data and integration with data visualization tools. To create an analysis, ORIO requires that each user create an email-associated account; the user receives notification of completed analyses through email messages. By default, analyses are private and only accessible by the creator; however, analyses may be made public such that anyone who has the URL could view the output (the examples presented below are public). ORIO allows for upload of user genomic feature lists through a web form and of user NGS data through HTTP. The web interface of ORIO is available at https://github.com/niehs/orio-web/. In addition to providing a manual and 'Getting Started' guide, its documentation includes instructions for deployment and creation of a development environment. Private ORIO deployments may be customized to specialized analytical approaches or experimental methods and be integrated into in-house NGS workflows.

Data visualization and statistical tests

ORIO provides tools for dynamic visualization of analysis results. The results of clustering approaches are presented in heatmaps annotated by tooltips. Raw read coverage values may be viewed using heatmaps and 2D scatterplots. Further investigation of read coverage may be performed through rank-order sorting of coverage heatmaps and quartiling of genomic features by independent data vectors. Non-parametric statistical tests may be applied to NGS and genomic feature groups for rigorous comparisons and hypothesis testing. Differences across quartiled genomic features are assessed using Anderso|Darling and Kruskal–Wallis tests. Differences across clustered genomic features are assessed using pairwise Mann–Whitney tests. These tests empower rigorous analysis of visualized differences and may imply functional association.

Data visualization by ORIO is implemented in Javascript and uses D3 (15). Statistical tests are implemented using NumPy and SciPy (13).

Hosted data

ORIO hosts 4506 human (hg19) and mouse (mm9) datasets from the first production run of the ENCODE project (6,16). When selecting hosted datasets to include in an analysis, ORIO implements dataset filters by metadata such as cell type, experiment type and antibody, allowing for rapid selection of relevant datasets. ORIO hosts recent assemblies for human (hg19 and hg38) and select model organisms (mouse: mm9 and mm10; fly: dm6). Sample genomic feature lists based on human and mouse RefSeq gene models are hosted (17). Additional annotations and assemblies can be easily added and will be included upon request.

Derivation of gene expression values

For some example analysis, additional supplemental analyses were performed. Gene expression values were derived from ENCODE RNA-seq datasets for mouse ES-Bruce4

Table 1. Benchmarking of ORIO runtimes with analyses of different sizes

		NGS datasets used					
		2	5	10	50	100	500
Feature list members	10	10	10	11	16	21	134
	50	11	24	12	16	23	156
	100	74	74	80	186	342	1668
	500	75	77	116	175	312	1929
	1000	75	77	90	176	339	1968
	5000	85	93	100	225	421	2541
	10 000	93	105	112	299	570	3348
	50 000	167	214	285	845	1614	9523
	100 000	243	346	568	1718	3237	19 702

Number of NGS datasets used and length of feature list are indicated in column and row headers, respectively. NGS datasets were selected at random from hosted ENCODE hg19 datasets. Feature lists were generated by selecting random positions in the human genome. Runtimes are reported in seconds of wall-clock time.

cells. Isoform expression values were found using Cuffnorm of the Cufflinks package (18) considering reported read alignments and RefSeq gene models (17). FPKM values for each isoform were averaged over two replicates. A single FPKM value was found for each gene by selecting the highest FPKM of any associated isoform.

Analysis of mouse embryonic stem cell transcription start sites

Published mouse embryonic stem cell Start-seq data (19) was used to call transcription start sites (TSSs). Prior to calling, reads with average Phred quality scores less than 20 were removed, and TruSeq adaptor sequences were trimmed using cutadapt (20). Reads were then aligned to the mm9 assembly using Bowtie (parameters: -m1 -v2 -X1000) (21). 5′ read end read coverages were then compiled into a bedGraph file and applied to TSS calling by Python script TSScall.py (default parameters). This calling approach is based on previously described methodologies (22,23). In short, TSSs called at single-nt positions with the most reads across bins spanning the entire genome. TSSs are only called at positions whose read count exceeds an FDR-based threshold (here, FDR = 0.001). TSSs were then grouped by distance; groups were made such that any two TSSs within 1000 bp of each other were placed in the same group. The TSS with the highest read count was taken forward as a representative TSS. TSScall.py and a script for selecting group representatives (TSSfilterToClusterRepresentatives.py) are available at https://github.com/niehs/TSScall/. The final TSS list used in analysis contained 163,274 TSSs. This number of TSSs is consistent with ranges reported in similar studies of genome-wide transcription in mammals (24).

Licensing

ORIO and ORIO-web are open source software released under the MIT License. License details can be found at https://github.com/NIEHS/orio/blob/master/LICENSE and https://github.com/NIEHS/orio-web/blob/master/LICENSE.

RESULTS AND DISCUSSION

An ORIO analysis begins with selection of a list of features at discrete genomic coordinates. A variety of different biological phenomenon may be represented as such a list. For instance, the activity of a transcription factor may be represented by protein binding sites determined empirically from ChIP-seq data (25,26) or predicted by occurrence of protein binding motifs (27,28). ORIO anchors its analysis to the selected feature list, effectively focusing on regions of the genome relevant to user interest.

ORIO iteratively calculates read coverage for each genomic feature and each NGS dataset (Figure 1A). Coverage values are found in user-defined windows anchored at each feature. Acknowledging that most NGS analytical methods involve some manipulation of read coverage, ORIO bases its analysis on coverage and is agnostic to the exact method used. Because read coverage is considered, ORIO may be applied to a wide variety of NGS data types, allowing for integration of different experiment types. Use of a relatively simple analytical approach was a deliberate choice that enhances ORIO's flexibility and makes its results readily interpretable. Even when using targeted analytical approaches specific to a given data type, read coverage is still informative. For instance, considerations specific to ChIP-seq data may be used in peak calling; however, results must often be validated through direct comparison to read coverage, either by eye or by performing meta-analyses of coverage values.

Coverage values are then used in integrative analysis. Both NGS datasets and genomic features are associated into informative groups by hierarchical and *k*-means clustering, respectively (Figure 1B and C). Analysis results are reported in dynamic displays alongside p-values from statistical tests. Clustering approaches and statistical tests were selected to minimize assumptions made about the analyzed data and broaden applicability. For instance, rank-order correlation and non-parametric statistical tests are applied due to the minimal assumptions made about the underlying data.

ORIO combines several informative analytical and visualization methods into a single tool, each tied to the execution of a single analysis run. We are unaware of another tool that combines the ability to rapidly compare NGS datasets with integrated statistical tests and clustering, though these functions are individually available (Table 2). With this and other features, ORIO differentiates itself based on ease of use and accessibility. ORIO achieves a high degree of flex-

ibility by supporting a diversity of data types, allowing incorporation of user-defined custom data, and by not requiring analysis on predefined features. All ORIO functions are accessible through a standalone web interface, with an intuitive data management system for adding and editing user data. Unique for NGS data visualization tools, ORIO hosts publicly available NGS datasets and annotations and performs computationally intensive analytical calculations on a remote server. By hosting 4506 human and mouse datasets from the ENCODE project (6,16), ORIO obviates the need for users to download and maintain prohibitively large data. Incorporating these files into ORIO enables filtering functions for rapid and facile selection of consortial data. Together with intuitive management of user-uploaded data and remote hosting of intensive analysis, users may readily contextualize their data amongst diverse datasets.

We have described the function of ORIO in general terms, but its applicability to biological questions may best be conveyed by example analyses. In the following examples, we apply ORIO analysis to (i) validate data derived from a ChIP-seq experiment, (ii) integrate ChIP-seq peaks with other NGS data, (iii) correlate gene expression with enrichment of histone modifications, (iv) recapitulate cell lineages using DNase-seq data and (v) characterize enhancer and promoter regions by a variety of NGS datasets (https://orio.niehs.nih.gov/quickstart/). We provide publicly accessible links to analysis results for each example.

Example 1: quality control by comparative validation of ChIP-seq data

NGS data quality is frequently first assessed using metrics such as read quality scores and duplication rates. However, even datasets with acceptable sequencing metrics can fail to provide biologically relevant information. Confidence in an experimental dataset may be improved by comparing it to validated experiments. This is especially important with techniques such as ChIP-seq where observed signal enrichment may be due to intrinsic variation found in input controls. ORIO complements common quality control procedures by allowing rapid comparison with validated datatsets similar in experimental design. We considered a publicly available H3K27ac ChIP-seq dataset from mouse ES-Bruce4 cells as an experimental test case and analyzed it against a diversity of other H3K27ac ChIP-seq datasets (16). Datasets from input controls were included in analysis as well as datasets from the mutually exclusive and functionally distinct H3K27me3 mark. A hosted list of 27 829 mouse TSSs from RefSeq was used as the ORIO feature list, restricting the analysis to biologically relevant promoter regions. Default parameter were used, defining analysis windows of 5 kb centered on each TSS. We present results from this analysis Figure 2; the analysis is publicly available at https://orio.niehs.nih.gov/dashboard/analysis/889-sample-validation-of-chip-seq-data/.

Hierarchical clustering show clear separate grouping of H3K27ac, H3K27me3, and input control datasets (Figure 2A). Replicates are tightly coupled in the resulting clustering. Importantly, test case data from ES-Bruce4 cells clusters well with other datasets, implying high data quality. Further supporting the separation between H3K27ac and controls, signal enrichment is coordinated across H3K27ac, H3K27me3, and input groups in feature clustering of RefSeq TSSs; an individual feature will generally have either high or low signal for a given group, but signal may not be consistently high across H3K27ac, H3K27me3, and input controls (Figure 2B). Notably, there is intrinsic variability in the input controls, which could easily be misinterpreted as signal in the experiment. Correlative analysis by ORIO allows the user to validate that differences in experimental signal are distinct from intrinsic variability in the input control. Due to the abundance of hosted data, ORIO can support rapid validation for many potential experimental datasets.

Example 2: integration of NGS data over ChIP-seq peaks

Peak calling is often applied to ChIP-seq datasets to define areas of enriched read coverage. We analyzed a list of 30 774 ChIP-seq peaks derived from the same ES-Bruce4 H3K27ac dataset validated in example 1 (16). These peaks correspond to regions in the genome enriched for the H3K27ac mark. We performed an ORIO analysis over these sites to see what data correlate with H3K27ac and to see if distinct functional states exist at H3K27ac-enriched regions. These functional states may display coincident genomic features such as histone modifications and transcription factor binding sites; in this way, though the analysis is centered on H3K27ac-enriched sites, other features are effectively investigated. We used the list of H3K27ac peaks as an ORIO feature list and then integrated all ENCODE ChIP-seq data from mouse Bruce4 and E14 ES cell lines, allowing exploration of transcription factors and other histone marks. Results from this analysis are presented in Figure 3 and are publicly available at https://orio.niehs.nih.gov/dashboard/analysis/850-sample-integration-over-chip-seq-peaks/.

H3K27ac is a mark associated with activated regions of the genome, including both promoters and enhancers. Therefore, H3K27ac marks coincide with other marks associated with promoters and enhancers, though promoters and enhancers themselves will be distinct in terms of chromatin state. ORIO provides informative clustering of ChIP-seq datasets at H3K27ac peaks; there are two distinct clusters of histone marks, one with promoter-associated marks and another with enhancer-associated marks (Figure 3A). A separate bar-plot view displays all correlations sorted by the absolute correlation value (Figure 3B). Though centered on H3K27ac peaks, ORIO analysis allows characterization of other genomic features. A relatively strong correlation is seen between H3K4me3 and H3K9ac marks. To allow more precise examination of pairwise relationships, ORIO creates rank-ordered heatmaps (Figure 3C), scatterplots (Figure 3D), and plots of read coverage quartiled by rank order (Figure 3E). Anderson–Darling and Kruskal–Wallis statistical tests are applied to rank-ordered quartiles of read coverage to evaluate the differences in quartile distributions. Data displays and statistical tests are complementary, allowing for in-depth characterization of biological phenomena and validation of initial observations. Here, the significant correlation between H3K4me3 and H3K9ac implies coincidence of these two marks in biologically active re-

Table 2. Comparison of ORIO features with similar tools

	Read coverage characterization over feature lists with one or more datasets			Statistical comparison of coverage values	Dataset clustering	Feature clustering	Feature to gene list annotation	Consortial data hosting	Accessible online
	Single dataset	Comparison with 2 datasets	Comparison with 3 or more						
ORIO	+	+	+	+	+	+	+	+	+
BioWardrobe (45)	+	+	-	+	-	-	+	-	+
ChIPseeker (46)	+	+	+	+	-	-	+	-	-
ChIPseeqer (47)	+	+	+	-	-	+	+	-	-
Deeptools (48)	+	+	+	-	+	+	-	-	*
Easeq (49)	+	+	+	-	+	+	+	+	-
GeneProf (50)	-	-	-	+	+	-	+	+	+
Homer (25)	+	+	+	-	-	-	+	-	*
ngs.plot (51)	+	+	+	-	-	+	-	-	*

'+' indicates that an identical or similar feature exists for a given tool. In the column 'Accessible online', '*' indicates that a tool is available through Galaxy (44); otherwise, '+' indicates that a tools is accessible by an independent web application.

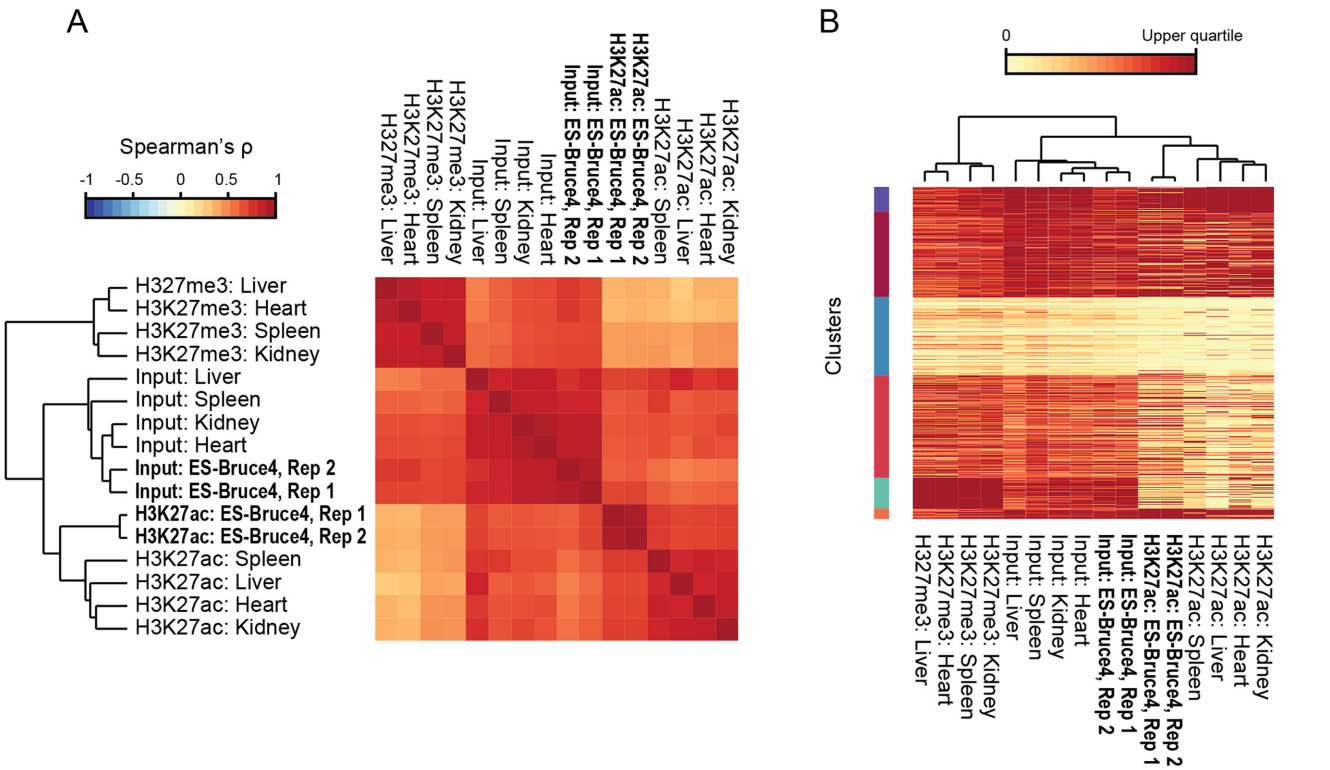

Figure 2. Example 1: Validation of test H3K27ac ChIP-seq data. (**A**) NGS dataset clustering. ORIO analysis was performed for ChIP-seq experiments from diverse tissues considering mouse RefSeq transcription start sites. H3K27ac datasets were run with input controls and the mutually exclusive H3K27me3 mark. Hierarchical clustering of samples is shown by dendrogram. (**B**) Clustering of transcription start sites by ChIP-seq data. Read coverage at transcription start sites is presented by heatmap where rows correspond to individual features and columns correspond to ChIP-seq datasets. Clusters (*k*-means; $k = 6$) are given on the left side of the plot. The dendrogram (top) reflects hierarchical clustering of ChIP-seq datasets shown in A. All plots were generated using ORIO.

gions, suggests their potential coordination in gene regulation, and recapitulates trends observed in heatmaps and scatterplots.

ORIO also clusters features based on their overlap with NGS datasets. Here, H3K27ac peaks are clustered by their coincidence with other histone marks. Clustering predominately separates peaks into those associated with promoters and/or enhancers; cluster 1 peaks, in regions with enhancer-associated marks like H3K4me1, are found in intergenic regions and gene bodies, while cluster 3 peaks, coincident with marks like H3K4me3, are found at promoters (Figure 3F). ORIO generates box plots and performs statistical tests that allow comparisons of individual clusters considering a given dataset; cluster 1 shows a significant enrichment of H3K4me1 signal over cluster 3, while cluster 3 shows a significant enrichment of H3K4me3 (Figure 3G

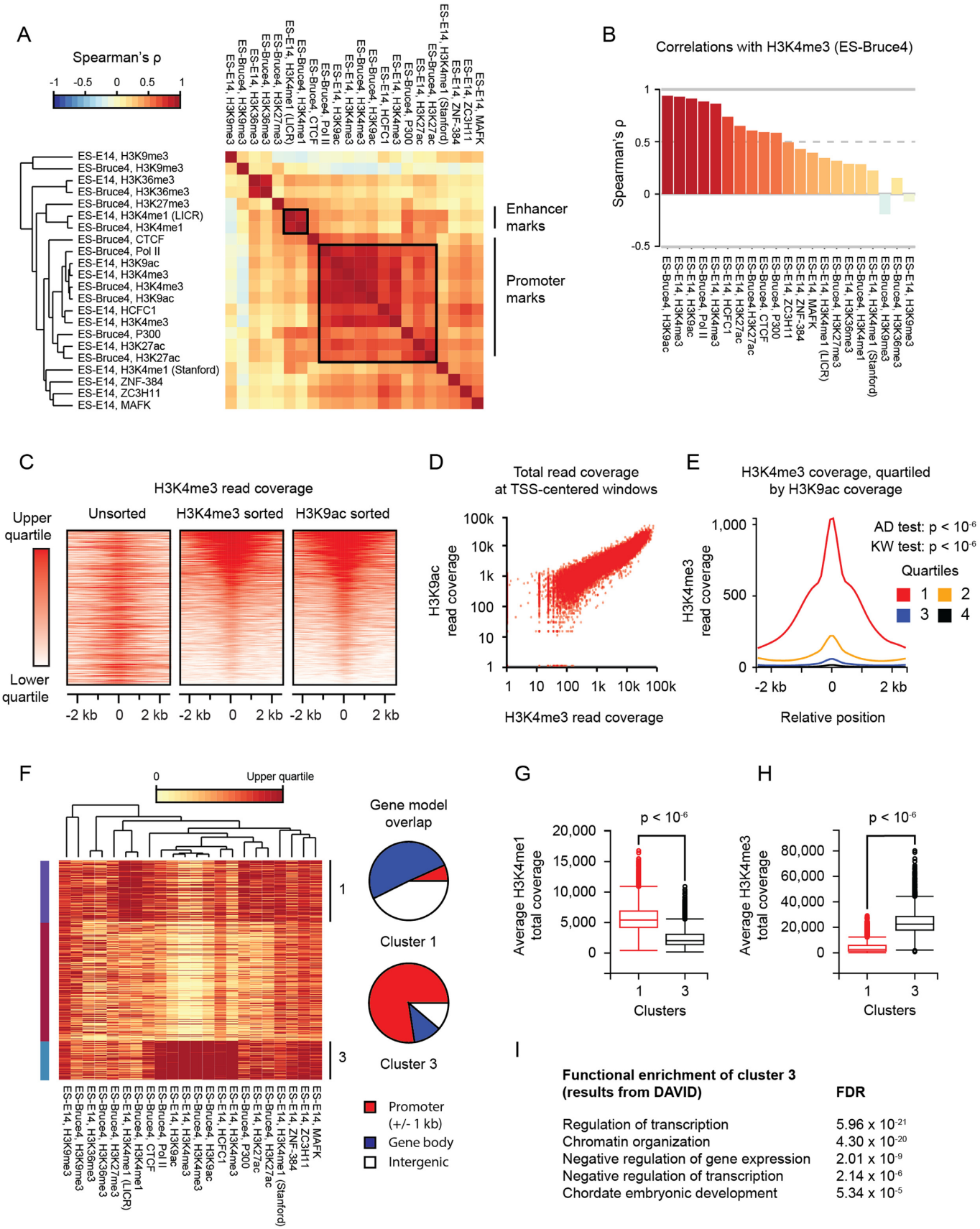

Figure 3. Example 2: ORIO analysis of H3K27ac ChIP-seq peaks. (**A**) NGS dataset clustering. ORIO analysis was performed for ES-Bruce4 and ES-E14 ChIP-seq experiments from ENCODE considering mouse ES-Bruce4 H3K27ac ChIP-seq peaks. Pairwise Spearman correlation values are displayed by

and H). ORIO automatically associates features with genes by proximity and allows these gene lists to be easily accessed in the user interface. These results are amenable for use in external tools such as gene ontology and functional annotation programs (29). Genes from cluster 3 show overlap with functions and processes associated with embryonic stem cells, such as chordate embryonic development and widespread negative transcription regulation commonly associated with a pluripotent state (Figure 3I). Clustering by ORIO recapitulates that H3K27ac is associated with distinct promoter and enhancer regions, each characterized by distinguishing genomic features (30).

Example 3: correlation of gene expression with enrichment of histone modifications

Integrating analysis of gene expression values with NGS data is a common way to investigate putative mechanisms of gene regulation. Determining the extent of gene expression from RNA-seq data typically requires analysis of read coverage while considering transcript models. ORIO provides no means of natively determining gene expression values. However, we anticipated that users will want to incorporate independently analyzed outside data into ORIO. To address this, ORIO allows flexible integration of non-NGS data through a generic data format. Users may provide individual values for each item in a genomic feature list in a tab-delimited text file. These values could be quantitative, such as gene expression values, or categorical, such as chromatin states. In ORIO, these data are referred to as a sort vector. When a sort vector is provided at the outset of an ORIO analysis, hierarchical clustering of NGS data is performed based on rank-order correlation of read coverage values with the sort vector. Sort vectors may be leveraged to allow creative integration of non-NGS data. Using a sort vector, we incorporated gene expression information into this example to find genomic features associated with active genes.

To quantify gene expression, FPKM values were calculated from an ENCODE RNA-seq dataset for mouse ES-Bruce4 cells. These FPKM values were then used as a sort vector in an ORIO analysis. To correlate gene expression with histone modification enrichment, the analysis was run with all ENCODE ChIP-seq datasets in ES-Bruce4 cells considering RefSeq TSSs as the feature list. Analysis was also run with the RNA-seq datasets used to calculate FPKM values. Results are presented in Figure 4 and are publicly available at https://orio.niehs.nih.gov/dashboard/analysis/853-sample-correlation-with-gene-expression/.

Considering genomic windows about features, ORIO presents sort vector correlations on a bin-by-bin basis (Figure 4A). As expected, RNA-seq read coverage values correlate well with expression values immediately downstream of TSSs. Here, RNA-seq read coverage represents steady-state levels of transcribed RNAs. Lower correlation values further downstream of TSSs likely reflect the lower coverage in downstream introns. Correlations also reveal histone marks associated with active promoters, such as H3K4me3. Correlations may be investigated at a given bin through scatterplot (Figure 4B) or for a given dataset by bar chart (Figure 4C). ORIO also allows for rank ordering of heatmaps and quartiling of read coverage by sort vector values. H3K4me3 heatmaps sorted by gene expression show a relationship between H3K4me3 read coverage and expression (Figure 4D), and Anderson–Darling and Kruskal–Wallis tests verify that the difference in the distributions of quartiles is significant (Figure 4E). Analysis by ORIO recapitulates that H3K4me3 modification is correlated with gene expression (31–33).

Example 4: recapitulation of cell lineages using DNase-seq

For the fourth and fifth examples, we applied ORIO to a comprehensive list of 163 274 TSSs derived from Start-seq data in mouse embryonic stem cells (19). ORIO allows custom-defined genomic features to center analysis. Here, features were determined by TSScall.py, a standalone Python script for calling TSSs from Start-seq data based upon previously described methods (23,34). Start-seq begins with the selection of short capped species from nuclear RNA isolates, ensuring that characterized transcripts are derived from early stages of transcription (22). The 5′ end of each Start-seq read corresponds with single-nucleotide resolution to the TSS of that transcript. Because transcripts are taken from the early stages of transcription, Start-seq describes TSSs of both mRNAs and short-lived non-coding RNAs, including RNAs produced at enhancers (35). As such, TSSs called from Start-seq describes active promoters and enhancers across the genome. TSSs were examined with all ENCODE DNase digital genomic footprinting (DGF) data available in mouse representing 22 different cell lines and tissues. DNase-DGF footprinting assays regions of the genome available for DNA-protein interactions and effectively characterizes the cis-regulatory framework of the target organism (36). The list of TSSs provides a catalog of

←

heatmap with hierarchical clustering of datasets indicated by the inset dendrogram. Groups of enhancer- and promoter-associated are indicated by boxes and by labels along the right side. (**B**) Boxplot of pairwise correlations (Spearman's ρ) with H3K4me3 ChIP-seq data. (**C**) Rank-ordered heatmaps of H3K4me3 read coverage over H3K27ac peaks. Heatmaps show read coverage from H3K4me3 ChIP-seq in ES-Bruce4 cells. From left to right, heatmap rows are unsorted, sorted by decreasing totals of H3K4me3 coverage, and sorted by decreasing totals of H3K9ac coverage. (**D**) Scatterplot ofH3K4me3 and H3K9ac read coverage over windows centered on H3K27ac peaks. (**E**) Plots of average H3K4me3 read coverage across genomic windows centered on H3K27ac peaks. Average coverages are quartiled by H3K9ac read coverage: the first quartile has the highest H3K9ac read coverage, and the fourth quartile has the lowest. *P*-values are found for Anderson–Darling (AD) and Kruskal–Wallis (KW) tests considering quartiled distributions of H3K4me3 read coverage. (**F**) Clustering of H3K27ac peaks as genomic features. Read coverage is displayed by heatmap, with each row corresponding to an individual feature and each column corresponding to an NGS dataset. NGS clustering is shown by dendrogram. Clusters (*k*-means; $k = 3$) are indicated along the left and right sides. Overlap of cluster features with RefSeq gene models is shown by pie charts. (**G** and **H**) Box plots of H3K4me1 and H3K4me3 read coverage in feature clusters 1 and 3. *P*-values were determined by Mann–Whitney test. (**I**) Gene ontology results for genes nearest H3K7ac peaks in cluster 3. Gene ontology analysis was performed by DAVID (29). All plots were generated using ORIO, with the except of pie charts in F; pie charts were generated using matplotlib.

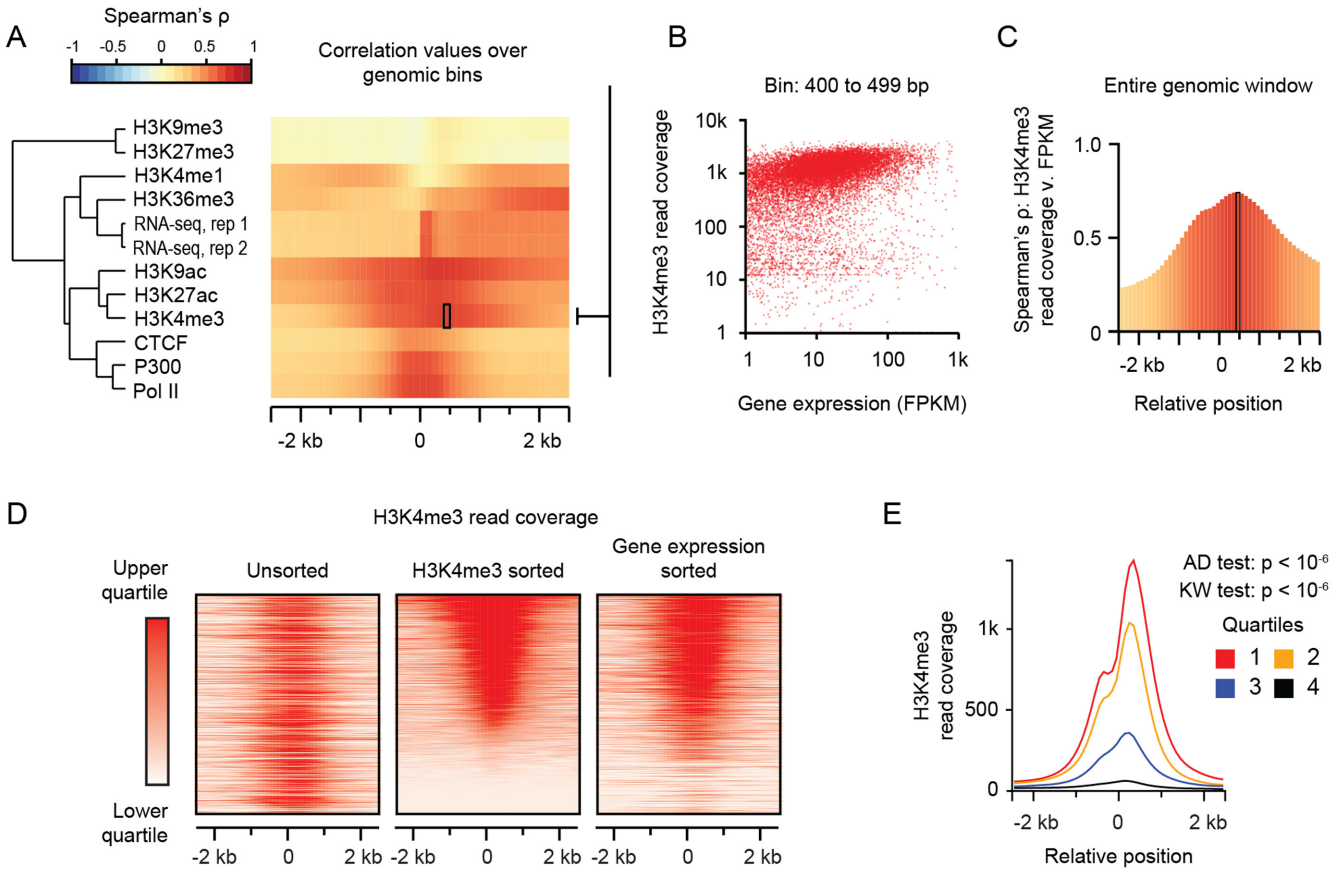

Figure 4. Example 3: ORIO analysis integrating gene expression with ChIP-seq and RNA-seq read coverage information. (**A**) Heatmap of correlations of read coverage values at transcription start sites with gene expression (FPKM). Each row corresponds to an individual NGS dataset, and each column corresponds to a genomic window bin. The left-side dendrogram was generated by hierarchical clustering of correlation values. (**B**) Scatterplot of gene expression values and H3K4me3 read coverage across transcription start sites in the 400–499 bp bin. This bin is indicated by the box in A. (**C**) Plot of correlation values across all window bins. Spearman correlation values were computed considering H3K4me3 read coverage at a RefSeq TSS and gene expression values (FPKM) at the nearest gene model. (**D**) Rank-ordered heatmaps of H3K4me3 read coverage. From left to right, heatmaps of H3K4me3 read coverage are unsorted, sorted by decreasing H3K4me3 read coverage, and sorted by decreasing gene expression of the closest gene as measured by FPKM. (**E**) Plots of average H3K4me3 read coverage for transcription start sites quartiled by gene expression. The first quartile corresponds to genes with the highest expression. Listed *P*-values are from Anderson–Darling (AD) and Kruskal–Wallis (KW) tests considering the quartiled distributions of total H3K4me3 read coverage. Plots were generated using ORIO.

enhancer and promoter regions under cis-regulatory control. Across different cells and tissues, many of these regions will be regulated in a similar fashion, but differences may be attributed in part to differences in cell lineage (37). We hoped to capture these differences in an ORIO analysis. Results from this example are presented in Figure 5 and are available at https://orio.niehs.nih.gov/dashboard/analysis/842-sample-recapitulation-of-cell-lineages/.

Considering the 22 samples, ORIO reveals that DNase-DGF data correlate well across all cell types, but higher correlation values are found when considering similar cell types (Figure 5). When ORIO analysis is applied to all mouse ENCODE DNase-DGF datasets across TSSs, clustering accurately recapitulates cell lineages. Clusters correspond to closely-related cells, such as lymphocytes, neuronal cells, and stem cells. Here, ORIO analysis of NGS data reflects underpinning biological differences in the cis-regulatory network between cells.

Example 5: characterization of enhancer and promoter regions by NGS datasets

ORIO was then applied to categorize TSSs into putative functional groups. Analysis was performed with all ENCODE datasets from Bruce4 C57BL/6 mouse embryonic stem cells (6) across mouse embryonic cell TSSs (19). These data were derived from RNA-seq and from ChIP-seq experiments targeting histone modifications and proteins widely involved in gene regulation, such as Pol II and CTCF. Results are presented in Figure 6 and are available at https://orio.niehs.nih.gov/dashboard/analysis/855-sample-characterization-of-enhancers-and-promoters/.

The TSSs described by Start-seq are found in active regions across the genome, including in regions with distinct functional roles, like promoters and enhancers. TSSs were clustered considering ChIP-seq and RNA-seq read coverage values (Figure 6A). *k*-means clustering results were examined with *k*-value of 10 to provide the greatest granularity allowed by ORIO. As evidenced by coverage val-

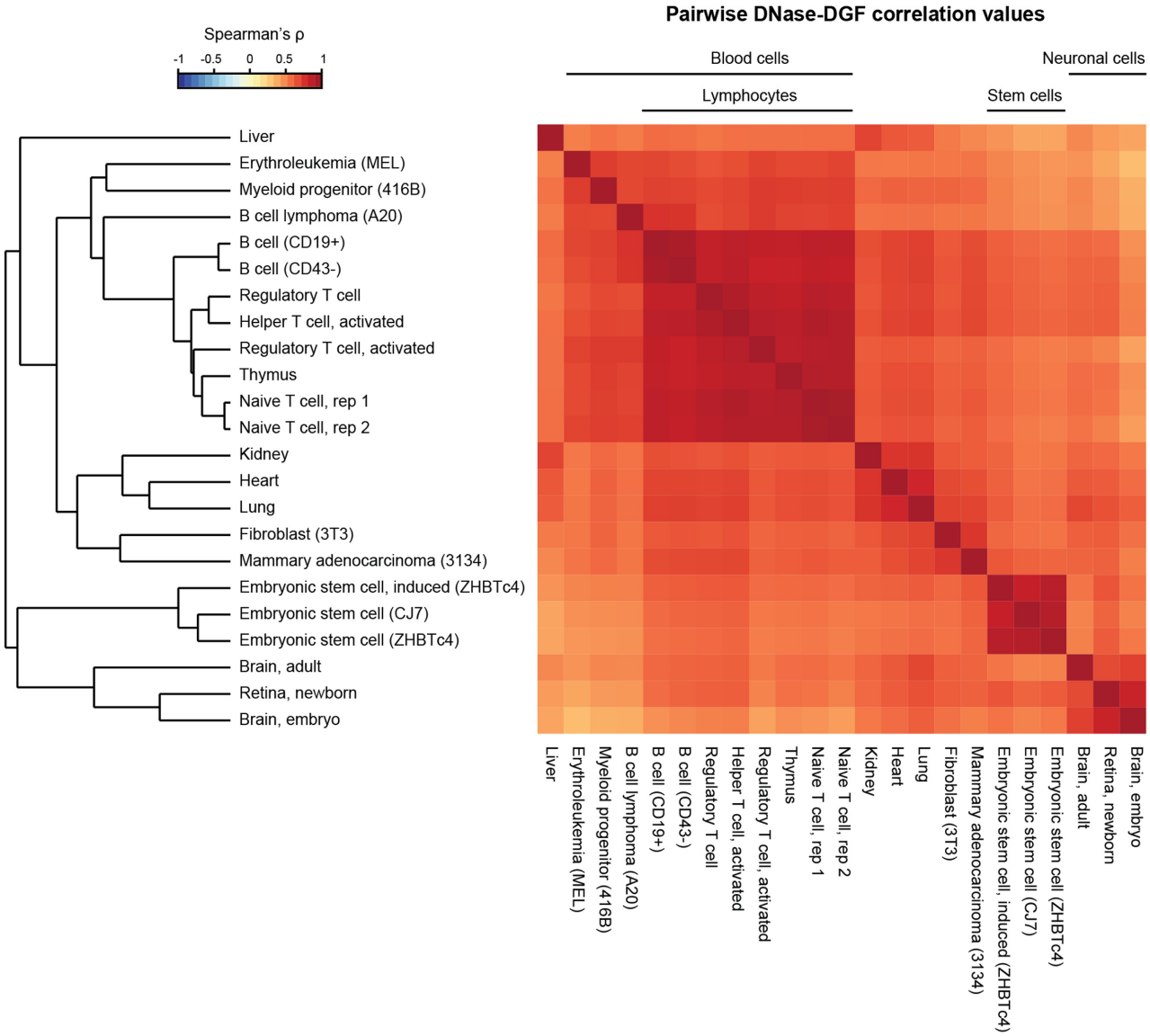

Figure 5. Example 4: Clustering of mouse DNase-DGF datasets by ORIO analysis. ENCODE DNase-DGF datasets from 22 cell and tissue types (6,36) were analyzed considering a feature list of 163,274 mouse TSSs. Correlation values (Spearman ρ) were found for each dataset pair and are displayed on a heatmap. Hierarchical clustering results were derived based on these correlation values and accurate recapitulate cell lineages. Clustering results are presented as a dendrogram, and cell subgroups are indicated along the top of the heatmap. Plots were generated using ORIO.

ues at cluster centroids (Figure 6B), clusters are enriched for histone modifications associated with distinct functional roles. These roles were confirmed by comparisons with RefSeq gene models (17) and chromatin states determined by chromHMM analysis (38,39) (Figure 6C). Clusters 2 and 7 show an enrichment of enhancer-associated marks like H327ac and H3K4me1 (40); consistent with this, most TSSs in this cluster display 'Strong Enhancer' chromatin states. Clusters 3 and 9 display an enrichment of H3K4me3, a mark associated with promoter regions (2,40). TSSs in this cluster are found predominantly in the promoter regions of RefSeq gene models. Cluster 3 shows strong overlap with 'Active Promoter' chromatin states. Cluster 9, with a distinct enrichment of repressive mark H3K27me3 (41), shows unique overlap with 'Poised Promoter' and 'Repressed' chromatin states; ORIO accurately segregates promoter-associated TSS clusters into functionally relevant groups. Though commonly used to interrogate gene expression, here RNA-seq provides a measure of transcription at a genomic locus. Cluster 10 shows enrichment for RNA-seq and H3K36me3 coverage (42); consistent with this, these TSSs occur predominantly within RefSeq gene models and in transcribed chromatin states. The remaining clusters 1, 4, 5, 6, and 8 all show relatively low signal with most histone modifications and are variably enriched for 'Heterochromatin' chromatin states. However, many TSSs in these clusters are found in enhancer- and transcribed region-associated chromatin states.

As evidenced by Example 5, ORIO informatively groups features into biologically relevant clusters. Coincident

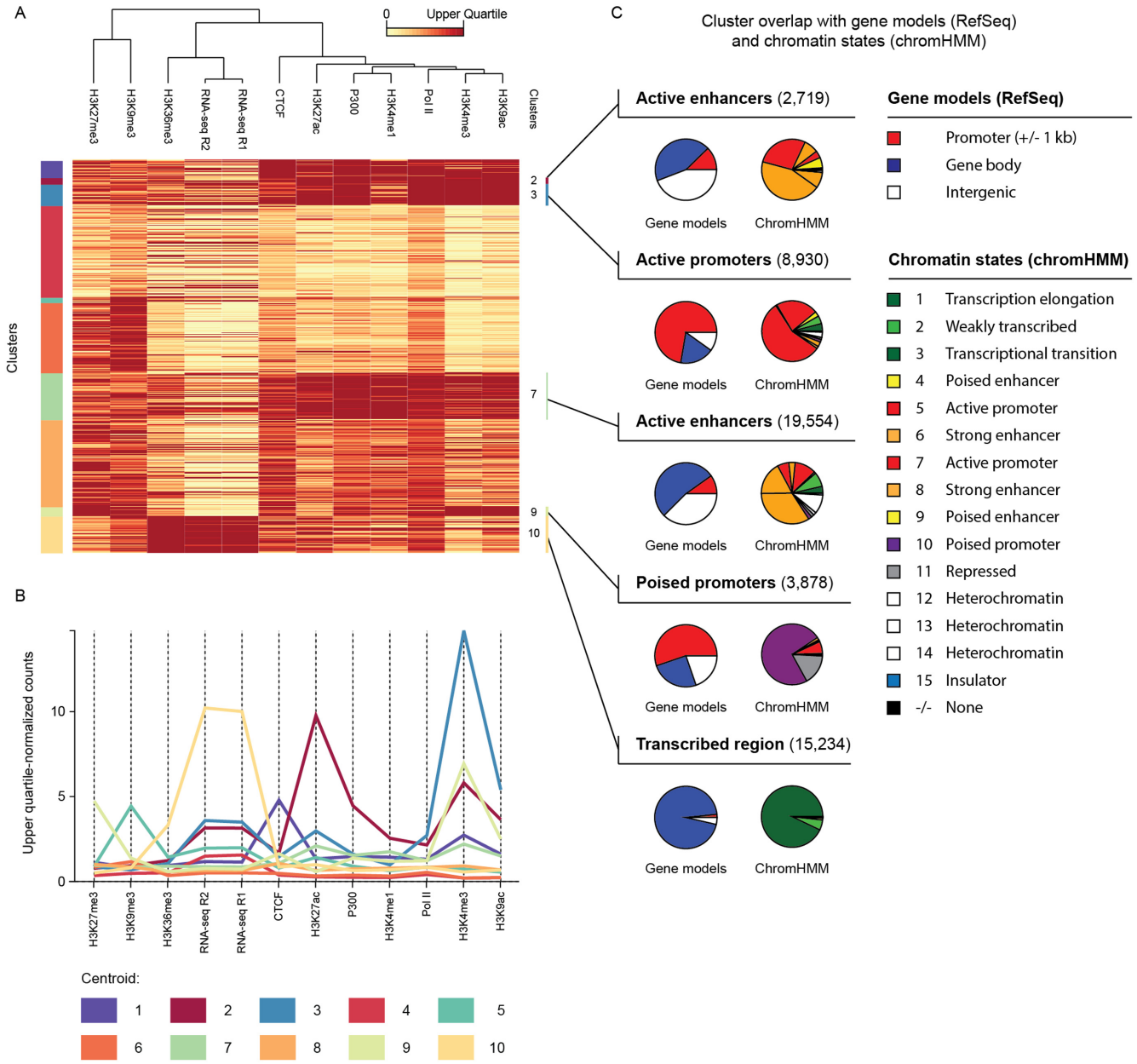

Figure 6. Example 5: ORIO analysis of mouse embryonic stem cell TSSs. (**A**) Clusters (k-means; $k = 10$) of TSSs called from embryonic stem cell Start-seq data. Clustering was performed considering observing read coverage of ENCODE Bruce4 C57BL/6 mouse embryonic stem cell datasets (6). Clustering results are displayed on an upper quartile-normalized heatmap where rows correspond to TSSs and columns to NGS datasets. Cluster membership is given by vertical bars on the left edge of the heatmap. (**B**) Normalized coverage values for the centroid of each cluster. Each column corresponds to an analyzed dataset. (**C**) Overlap of each cluster with RefSeq gene models (17) and chromHMM chromatin states (38,39). Overlap is displayed using pie charts. The redundancy of chromHMM state names is characteristic of this approach with labels such as 'strong enhancer' and 'heterochromatin' being applied to multiple states. Clustering heatmap and centroid plot were generated using ORIO; pie charts were generated using matplotlib (43).

marks at individual clusters also imply coordinated function. For instance, cluster 9 implies interplay between H3K4me3 and H3K27me3, two histone modifications associated with opposing activation and repression functions. Cluster membership provides a list of many genomic loci where this association may be investigated in detail.

CONCLUSION

ORIO provides a framework for rapid and informative integration of NGS data. As shown by numerous examples, ORIO performs applications ranging from data validation to hypothesis-driven investigations and provides biologically relevant results that recapitulate independent analyses. An intuitive web interface and extensive help documentation makes ORIO accessible to life scientists with minimal computational expertise. ORIO's acceptance of widely used

data formats makes it amenable to integration with other bioinformatics tools, such as peak callers, and its modern architecture makes it suitable to outside development toward specialized application. ORIO will facilitate genome-wide exploration of NGS data by life scientists and will support scientists transition between discovery, hypothesis generation, and more detailed investigations through robust analysis, informative visualization, and rigorous statistical tests.

ACKNOWLEDGEMENTS

We gratefully acknowledge Frank S. Day for support with computational infrastructure. Guidance from Dan Gilchrist helped steer early planning and design. Ben Scruggs and James Ward provided thoughtful insight during tool development. We thank Paul Wade, Guang Hu, and Jason Li for critical comments during review of the manuscript.

FUNDING

Intramural Research Program of the National Institute of Environmental Health Sciences, NIH [Z01 ES103312 to D.C.F. and Z01 ES101987 to K.A.]. Funding for open access charge: National Institute of Environmental Health Sciences [Z01 ES103312 to D.C.F.].
Conflict of interest statement. None declared.

REFERENCES

1. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 53–59.
2. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 311–322.
3. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 1213–1218.
4. Johnson,D.S., Mortazavi,A., Myers,R.M. and Wold,B. (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.
5. Wang,Z., Gerstein,M. and Snyder,M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.
6. Encode Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
7. de Dieuleveult,M., Yen,K., Hmitou,I., Depaux,A., Boussouar,F., Bou Dargham,D., Jounier,S., Humbertclaude,H., Ribierre,F., Baulard,C. *et al.* (2016) Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. *Nature*, **530**, 113–116.
8. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
9. Thorvaldsdottir,H., Robinson,J.T. and Mesirov,J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.*, **14**, 178–192.
10. Lavender,C.A., Cannady,K.R., Hoffman,J.A., Trotter,K.W., Gilchrist,D.A., Bennett,B.D., Burkholder,A.B., Burd,C.J., Fargo,D.C. and Archer,T.K. (2016) Downstream antisense transcription predicts genomic features that define the specific chromatin environment at mammalian promoters. *PLoS Genet.*, **12**, e1006224.
11. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
12. Kent,W.J., Zweig,A.S., Barber,G., Hinrichs,A.S. and Karolchik,D. (2010) BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, **26**, 2204–2207.
13. van der Walt,S., Colbert,S. and Varoquaux,G. (2011) The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30
14. Django Core Team. (2016) *Django: A Web Framework for the Python Programming Language*. Django Software Foundation, Lawrence, Kansas.
15. Bostock,M., Ogievetsky,V. and Heer,J. (2011) D3: data-driven documents. *IEEE Trans. Vis. Comput. Graph.*
16. Mouse Encode Consortium, Stamatoyannopoulos,J.A., Snyder,M., Hardison,R., Ren,B., Gingeras,T., Gilbert,D.M., Groudine,M., Bender,M., Kaul,R. *et al.* (2012) An encyclopedia of mouse DNA elements (mouse ENCODE). *Genome Biol.*, **13**, 418.
17. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
18. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
19. Williams,L.H., Fromm,G., Gokey,N.G., Henriques,T., Muse,G.W., Burkholder,A., Fargo,D.C., Hu,G. and Adelman,K. (2015) Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. *Mol. Cell*, **58**, 311–322.
20. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
21. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
22. Nechaev,S., Fargo,D.C., dos Santos,G., Liu,L., Gao,Y. and Adelman,K. (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, **327**, 335–338.
23. Scruggs,B.S., Gilchrist,D.A., Nechaev,S., Muse,G.W., Burkholder,A., Fargo,D.C. and Adelman,K. (2015) Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Mol. Cell*, **58**, 1101–1112.
24. Core,L.J., Martins,A.L., Danko,C.G., Waters,C.T., Siepel,A. and Lis,J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311–1320.
25. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
26. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
27. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
28. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
29. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.
30. Creyghton,M.P., Cheng,A.W., Welstead,G.G., Kooistra,T., Carey,B.W., Steine,E.J., Hanna,J., Lodato,M.A., Frampton,G.M., Sharp,P.A. *et al.* (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 21931–21936.
31. Araki,Y., Wang,Z., Zang,C., Wood,W.H. 3rd, Schones,D., Cui,K., Roh,T.Y., Lhotsky,B., Wersto,R.P., Peng,W. *et al.* (2009) Genome-wide analysis of histone methylation reveals chromatin state-based regulation of gene transcription and function of memory CD8+ T cells. *Immunity*, **30**, 912–925.

32. Karlic,R., Chung,H.R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 2926–2931.
33. Rahl,P.B., Lin,C.Y., Seila,A.C., Flynn,R.A., McCuine,S., Burge,C.B., Sharp,P.A. and Young,R.A. (2010) c-Myc regulates transcriptional pause release. *Cell*, **141**, 432–445.
34. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
35. Kim,T.K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K., Kuersten,S. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.
36. Hesselberth,J.R., Chen,X., Zhang,Z., Sabo,P.J., Sandstrom,R., Reynolds,A.P., Thurman,R.E., Neph,S., Kuehn,M.S., Noble,W.S. *et al.* (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.
37. Andersson,R., Gebhard,C., Miguel-Escalada,I., Hoof,I., Bornholdt,J., Boyd,M., Chen,Y., Zhao,X., Schmidl,C., Suzuki,T. *et al.* (2014) An atlas of active enhancers across human cell types and tissues. *Nature*, **507**, 455–461.
38. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
39. Bogu,G.K., Vizan,P., Stanton,L.W., Beato,M., Di Croce,L. and Marti-Renom,M.A. (2016) Chromatin and RNA Maps Reveal Regulatory Long Noncoding RNAs in Mouse. *Mol. Cell. Biol.*, **36**, 809–819.
40. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
41. Boyer,L.A., Plath,K., Zeitlinger,J., Brambrink,T., Medeiros,L.A., Lee,T.I., Levine,S.S., Wernig,M., Tajonar,A., Ray,M.K. *et al.* (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, **441**, 349–353.
42. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T.K., Koche,R.P. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
43. Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
44. Afgan,E., Baker,D., van den Beek,M., Blankenberg,D., Bouvier,D., Cech,M., Chilton,J., Clements,D., Coraor,N., Eberhard,C. *et al.* (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, **44**, W3–W10.
45. Kartashov,A.V. and Barski,A. (2015) BioWardrobe: an integrated platform for analysis of epigenomics and transcriptomics data. *Genome Biol.*, **16**, 158.
46. Yu,G., Wang,L.G. and He,Q.Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.
47. Giannopoulou,E.G. and Elemento,O. (2011) An integrated ChIP-seq analysis platform with customizable workflows. *BMC Bioinformatics*, **12**, 277.
48. Ramirez,F., Dundar,F., Diehl,S., Gruning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–191.
49. Lerdrup,M., Johansen,J.V., Agrawal-Singh,S. and Hansen,K. (2016) An interactive environment for agile analysis and visualization of ChIP-sequencing data. *Nat. Struct. Mol. Biol.*, **23**, 349–357.
50. Halbritter,F., Vaidya,H.J. and Tomlinson,S.R. (2011) GeneProf: analysis of high-throughput sequencing experiments. *Nat. Methods*, **9**, 7–8.
51. Shen,L., Shao,N., Liu,X. and Nestler,E. (2014) ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, **15**, 284.